Kevin Buckley (Newcastle University)

UNCOVERING LINGUISTIC LINEAGE THROUGH USING A CHARACTER N-GRAM BASED DIALECT CLASSIFIER

Quantitative approaches to analysing diachronic change have become popular in examining historical languages (Piotrowski 2012). Character N-grams, which are N-sized letter collocations, are a well-used method in analysing written text (Cavnar & Trenkle 1994) and have been used in language classification, such as in the program Textcat (Feinerer et al. 2013). Buckley and Vogel (2019) used profiles of character N- grams to examine change in historical English over time and to detect features of change across historical epochs.

This paper attempts to create a dialect classifier based on character N-gram features. The chosen linguistic context is Middle English (ME), using the Middle English Grammar Corpus (Stenroos et al. 2011). The classifier successfully delineates between a cluster of Northern regions and a cluster of Midlands and Southern areas.

Using this functioning text classifier, ME texts can be positioned in the regions they are most similar to. The analyses showed that texts of known origin are placed in a correct dialect cluster with high accuracy. Furthermore texts of Older Scots can be classified in relation to these ME regional clusters. Through this a quantitative confirmation that Older Scots is closest to Northern ME was found and that Older Scots texts are most similar to the northernmost counties of the Northern cluster.

Scots texts are analysed over time using the Helsinki Corpus of Older Scots (Meurman-Solin 1995) and it was found that texts become less close to Northern ME texts over time, signalling language change away from medieval forms. The features shared between Scots and Northern ME can be abstracted showing that this method can be a useful tool in examining the relationship between varieties of a language and the strength of the relationship between varieties over time.

References

Buckley, Kevin and Carl Vogel. 2019. "Using Character n-Grams to Explore Diachronic Change in Medieval English". *Folia Linguistica* 53.s40-2: 249–299.

Cavnar, William B. and John M. Trenkle. 1994. "n-Gram-Based Text Categorization". In: *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval* (Vol. 161175).

Feinerer, Ingo, Christian Buchta, Wilhelm Geiger, Johannes Rauch, Patrick Mair, and Kurt Hornik. 2013. "The textcat Package for n-Gram Based Text Categorization in R". *Journal of Statistical Software* 52.6: 1–17.

Meurman-Solin, Anneli. 1995. "A New Tool: The Helsinki Corpus of Older Scots (1450–1700)". *ICAME Journal* 19*:* 49-62.

Piotrowski, Michael. 2012. *Natural Language Processing for Historical Texts*. San Rafael, CA: Morgan & Claypool Publishers.

Stenroos, Merja, Martti Mäkinen, Simon Horobin, and Jeremy Smith. 2011. *The Middle English Grammar Corpus,* version 2011.1. University of Stavanger.